RECEIVED
CENTRAL FAX CENTER
MAR 17 2006

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BOARD OF PATENT APPEALS AND INTERFERENCES

In re patent application of:

Kreulen, et al.

Serial No.: 09/629,831

Filed: July 31, 2000

Group Art Unit: 2176

Examiner: Smith, Peter J.

Atty. Docket No.: AM9-99-0157

Certificate of Transmission by Facsimile

I hereby certify that this correspondence is being facsimile transmitted to the United States Patent and Trademark Office (Fax No. 571-273-8300) on 03/17/06

Frederick W. Gibb, III

For: METHOD FOR GENERATION OF AN N-WORD PHRASE DICTIONARY FROM A TEXT CORPUS

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPELLANTS' REPLY BRIEF

In response to the Examiner's Answer dated January 19, 2006, Appellants submit the following comments.

As explained in greater detail in Appellants' Appeal Brief (for example, section V, Summary, beginning on page 2) the claimed invention generally relates to the analysis of unstructured text documents which, by analyzing such items as frequency of the occurrence various terms/phrases within the text documents, can automatically create classifications for the various documents, and then sort the documents into the different automatically created classifications. Therefore, the claimed invention can be considered a form of data mining. One of the problems with analyzing text documents relates to the amount of data that must be processed, and there is always a problem balancing the trade-off of computer resources (hardware, computational speed, etc.) against accuracy and/or complexity of the document analysis.

While many conventional solutions concentrate on removing common terms from the dictionary (removing punctuation, replacing words with synonyms, removing stop words, removing duplicates words, clustering, etc.) in order to utilize less computer resources, the claimed invention takes a completely different approach and allows the automated methodology to decide which terms to remove. Thus, while conventional solutions require the manual creation of lists of words to be removed from the document analysis, the claimed invention simply allows the user to enter the maximum dictionary size. From the maximum dictionary size defined by the user, the automated analysis only includes the most frequently used words that the maximum dictionary size will accommodate into the document analysis process. Therefore, by simply inputting a maximum dictionary size, the claimed invention automatically removes the less frequently used words from the document analysis, thereby maximizing the use of the limited computer resources.

While the claimed invention can be used in combination with conventional methodologies that remove punctuation, minimize synonyms, remove stop words, remove duplicate words, etc., the claimed invention goes beyond such methodologies by providing the user with a means to control the balance and tradeoff between computer resources and accuracy/complexity by adjusting the maximum dictionary size. Thus, by using the "maximum dictionary size" as the vehicle to control how many terms are to be used in the phrase search (e.g., limiting the size of the dictionary before the frequency of phrases in the document that contain words in the dictionary is determined), the invention provides an automated methodology which, without additional user input, reduces the size of the data that must be processed.

Of all the applied prior art references, Kostoff is the only reference proffered as teaching the core of the claimed invention. The secondary references Kirsch, Kobayashi, and Turney are only presented for teaching concepts such as removing punctuation, replacing words with synonyms, removing stop words, removing duplicates words, clustering, etc. (see page 8, 3rd paragraph of the Examiner's Answer). However, Kostoff does not teach any form of limiting or adjusting the maximum dictionary size, but instead

only teaches removing a manually created list of trivial phrases from the dictionary before searching the associated documents. This appears to be a fundamental flaw in the rejection, because the rejection relies upon this teaching in Kostoff for the conclusion that it would have been obvious to limit/adjust the maximum dictionary size.

For example, the bottom of page 3 and top of page 4 of the Examiner's Answer states that Kostoff does not specifically teach inputting a maximum dictionary size and limiting the dictionary to the inputting maximum dictionary size, such that the dictionary contains less than all words in the documents. However, Kostoff does acknowledge the importance and limitation of memory size for sorting a list of trivial words. Similarly, see the discussion in section 10 of the Examiner's Answer. Such reasoning is clearly based upon hindsight because nothing within Kostoff contains any teaching or any suggestion that the maximum dictionary size should be altered. Instead, Kostoff merely states the well-known fact that is important to maximize the utilization of limited computer resources, and in order to do so, Kostoff suggests removing a manually-created list of trivial words from the dictionary before searching the text documents. There is nothing within such a teaching which would motivate one ordinarily skilled in the art to limit the maximum dictionary size so as to restrict the number of words that are utilized to determine the frequency of phrases within the document analysis process (and thereby maximize the utilization of the limited computer resources). Thus, because of this gap between what is claimed and what is taught by the prior art of record, Appellants submit that the Examiner has engaged in hindsight reasoning and has suggested a modification to Kostoff that is based, solely upon Appellants' disclosure. Such reasoning is improper and is grounds to remove the rejections herein.

At the top of page 4, the Examiner's Answer confuses the analysis by suggesting that the selection of often repeated phrases of high user interest (referred to as pervasive theme areas (PTA's)) can provide some motivation for limiting the size of the dictionary of words. As suggested in col. 1, lines 20-36 of Kostoff PTA's can be utilized to direct and focus the indexing process that is performed to produce more meaningful results. The discussion of PTA's within Kostoff has nothing whatsoever to do with limiting the

size of the dictionary of words because the PTA analysis can only occur after the dictionary of words is established and the full dictionary of words is utilized in Kostoff to search for phrases. More specifically, the claimed invention creates a dictionary of the most frequently occurring words as limited by the maximum dictionary size and then (after creating the dictionary or words) determines the frequency of phrases that utilize such words within the database of text documents (claim 1). To the contrary, as shown by tables 1 and 2 (appearing between col. 5-6 of Kostoff) the applied prior art reference Kostoff utilizes the full word dictionary (as possibly reduced by the removal of trivial words/phrases) in order to search for single, double, and triple would phrases (Table 1) in order to present the user with various phrases that may be of interest after which the user can provide some guidance to have the indexing of the documents within the database in a PTA process (Table 2).

Thus, the PTA processing described in Kostoff only begins after the full un-truncated dictionary of words has been used to identify single, double, and triple word phrases. Quite to the contrary, with the claimed invention, before the dictionary is allowed to be used to search for phrases, it is limited (and less frequently used words are removed) by the maximum dictionary size. In other words, all discussions relating to the PTA analysis/process in Kostoff (col. 5-12) occur after the full (un-truncated) dictionary of words has been utilized to identify phrases. To the contrary, the claimed invention truncates the dictionary of words before it is used to search for phrases. Thus, the comments regarding PTA analysis within Kostoff in the Examiner's Answer have nothing whatsoever to do with the claimed process of limiting the maximum dictionary size to remove words from the full dictionary (before performing the process of searching for phrases and determining the frequency of phrases within the text documents in the database). Instead, the reference to the PTA analysis section of Kostoff within the Examiner's Answer serves to confuse the reader and does not provide the reader with a true and complete understanding of the teachings of Kostoff as they relate to the dictionary of words that is used to search for phrases within the text documents in the database, as claimed.

To the contrary, as explained in col. 4, lines 39-55 of Kostoff, only the trivial phrases are removed immediately prior to processing and "the system and methodology are required to use the entire full-text database to create lists of phrases" during the identification of PTA's (col. 4, lines 50-55). In other words, the only hint of alteration of the dictionary of words (prior to searching the documents for phrases containing such words) in Kostoff relates to removal of the manually-created list of trivial phrases, and all discussion relating to pervasive theme areas (PTAs) in col. 5-12 of Kostoff relates only to the manner in which the phrases will be sorted, which is irrelevant to the claimed invention.

As described in the last paragraph in col. 4 (lines 50-68) the full dictionary of words (possibly with trivial words removed) is utilized to search the text documents to locate various phrases. After this searching process, the user identified preferences that are discussed in col. 5-12 are utilized to prefer certain phrases over others, and thereby focus the results produced by the document analysis. Thus, because the full dictionary of words is used to locate the phrases before they are altered by any form of PTA analysis, all discussions within Kostoff relating to PTA's do not alter the dictionary of words in any manner whatsoever, and the obviousness conclusion drawn in the Examiner's Answer which relies upon the teachings of PTA's in col. 5-6 of Kostoff is incorrect and confusing. As shown above, the only alteration to the dictionary of words prior to utilizing the dictionary of words to search for phrases in the text documents taught by Kostoff is the removal of the list of manually-created trivial phrases. The entire PTA analysis discussion within Kostoff occurs after the full dictionary of words has been utilized to search for phrases within the database of text documents. Therefore, it is Appellants' position that the references to the teachings regarding PTA's in Kostoff (for example, col. 5-6) have nothing whatsoever to do with adjusting the size of the dictionary of words, and that such references in the Examiner's Answer merely serve to confuse the reader regarding the true teachings of Kostoff.

With respect more specifically to the claimed invention, referring to Figure 1, the invention performs a "first pass" (independent claim 6) on the set of text documents, as

shown in the item 10. Next, in item 11, the invention creates a Hashtable and keeps only the most frequently occurring words in the Hashtable. Thus, the invention finds the V most frequently occurring words in the word-count Hashtable and conserves memory by removing from the Hashtable all words that occur with less frequency than the V most frequently occurring words. This is defined in the independent claims as "determining frequency of each word in each of said documents; creating a dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size, such that said dictionary contains less than all words in said documents."

As described on page 15, lines 1-9 of the application, previous methods for generating a dictionary from a text corpus focused on individual words only or have generated phrases based on a linguistic analysis. The invention's methodology is purely lexical in nature and thus generalizes to multiple languages and to ungrammatical text. Previous methodologies have suggested a lexical phrase generation technique and have not described the space and time efficient implementation for discovering such phrases that the invention utilizes. The invention's implementation is designed to quickly find a maximal frequency term dictionary of a given size using the smallest possible amount of memory.

It is Appellants' position that Kostoff does not teach one ordinarily skilled in the art to limit which words can be added to the dictionary according to the "maximum dictionary size." Independent claims 1 and 11 provide for "creating a dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size." Therefore, with the invention, the decision of which words to include in, or exclude from the dictionary is determined just by entering the "maximum dictionary size". To the contrary, with Kostoff, the manually created list of "trivial" words that are excluded from the dictionary is used to limit which words are excluded from the dictionary (col. 4, lines 39-42).

Contrary to the highly manual process described in Kostoff, the claimed methodology is fully automated (the only input required being the "maximum dictionary size", which can simply be equal to the available memory or manually preset by the user),

while Kostoff requires the user to manually create the trivial phrase list (col. 4, lines 39-42). The efficiency gains of the automated inventive methodology when compared to the manual system described in Kostoff are substantial.

Further, the removal of trivial words ("to", "if", etc.) in Kostoff is actually more similar to the claimed removal of a manually created list of "stop" words (the, and, a, there, is, than) as defined by dependent claims 2-3, 7-8, and 12-13. The rules of claim differentiation and construction provide that each claim in a patent is presumptively different in scope. Therefore, the removal of trivial stop words in the dependent claims is different that the removal of words based on the maximum dictionary size in the independent claims. Here, the removal of a manually created list of trivial phrases ("to", "if", etc.) in Kostoff is equivalent to the claimed removal of a manually created list of stop words (the, and, a, there, is, than). Thus, the claimed method of limiting the dictionary according to a maximum size is a distinct feature from the removal of trivial or stop words and phrases. Therefore, it is Appellants' position that the discussion in Kostoff regarding the list of trivial words and phrases teaches no more than what is performed when the claimed invention removes stop words. There is nothing with Kostoff which would suggest that this removal of trivial or stop words would lead one ordinarily skilled in the art to limit which words are to be included in the dictionary according to a "maximum dictionary size".

The creation of a manual list of trivial words ("to", "if", etc.) and its removal from the dictionary does not suggest the claimed automated methodology which simply and automatically limits the dictionary using a size limit. It is Appellants' position that the requirement that a manually created list be used to limit the dictionary size teaches away from the claimed automated methodology, which does not require the user to specify any words, but instead, merely eliminates the least frequent words from the dictionary. Further, the claimed invention may actually include all "trivial" words (if these stop words are not otherwise removed as provided in the dependent claims) as these words may be the most common. Again, the claimed invention removes the "most frequently occurring words in said documents as limited by said "maximum dictionary size" and

trivial or stop words may actually be the most common (if otherwise not removed in a separate processing step).

One difference between the claimed invention and Kostoff is that the size of the dictionary of words is limited before the frequency of phrases in the document that contain words in the dictionary is determined. This is important because the number of phrases grows exponentially with the size of the corpus. Simply removing a list of trivial phrases may not reduce the dictionary size (especially if the manually created list of trivial phrases finds no matches in the dictionary). By reducing the size of the dictionary before determining the frequency of phrases containing words in the dictionary, the claimed invention produces exponential gains in processing speed and memory usage. In other words, the claimed invention involves more than just reducing the dictionary to meet a memory constraint. In the claimed invention, the dictionary is reduced at a point in the processing that allows the method to substantially simplify the subsequent process of determining the frequency of phrases in the document containing words in the dictionary.

The claimed invention first limits the dictionary to only the top number of most frequently occurring words and then "after creating said dictionary" (claims 1 and 11) only considers phrases that contain these words. The invention avoids maintaining a list of all potential phrases in the text corpus. The problem with maintaining all potential phrases is that the number of phrases grows exponentially with the size of the corpus. The invention avoids this problem by fixing the size of the dictionary up front (user specified "maximum dictionary size", M), then finding the M most frequent words, and then only creating phrases using these M most frequent words. To the contrary, the Kostoff patent creates a list of potentially all words and N-word phrases sorted by frequency. This is not practical for a large text corpus since such a list would be too large for most computer memories to hold.

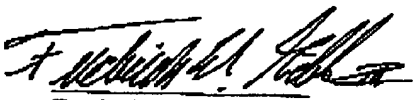
Thus, the claimed invention can search the associated document for phrases that contain only these terms and produce a dictionary of most frequently occurring phases and terms. By using the "maximum dictionary size" as the vehicle to control how many

terms are to be used in the phrase search (e.g., limiting the size of the dictionary before the frequency of phrases in the document that contain words in the dictionary is determined), the invention provides an automated methodology which, without additional user input, reduces the size of the data that must be processed.

In view the forgoing, the Board is respectfully requested to reconsider and withdraw the rejections of claims 1-17.

Please charge any deficiencies and credit any overpayments to Attorney's Deposit Account Number 09-0441.

Respectfully submitted,



Frederick W. Gibb, III
Registration No. 37,629

Date: 03/17/06
Gibb I.P. Law Firm, LLC
2568-A Riva Road, Suite 304
Annapolis, MD, 21401
301-261-8071
Customer No. 29154